

Prediction of Social Media Consumer Behaviour Using Machine Learning

Supervised by Dr. Priyanka Shukla

M. Sajid, P. Kumar, P. Ray, S. Majhi, S. Das, V. Kumar

Department of Mathematics, Indian Institute of Technology Madras, 600036, India

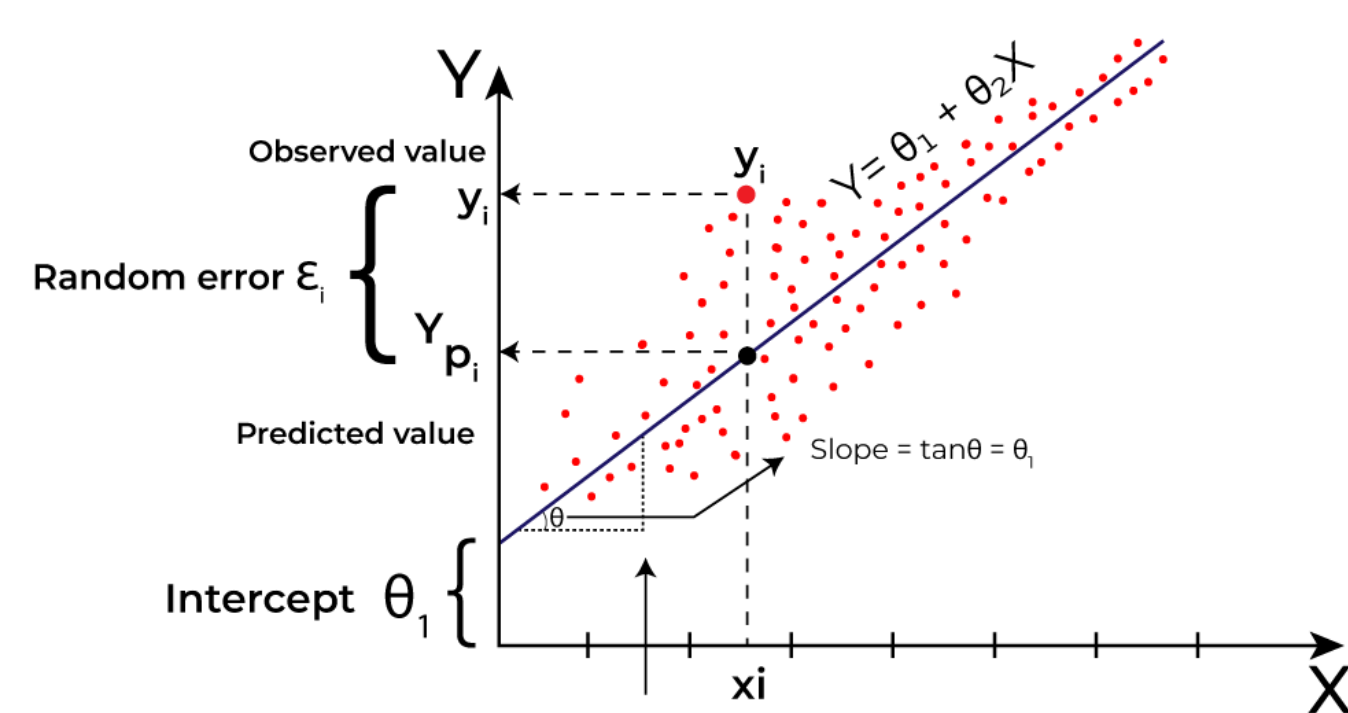
Motivation

Why do we need to predict social media behaviour?

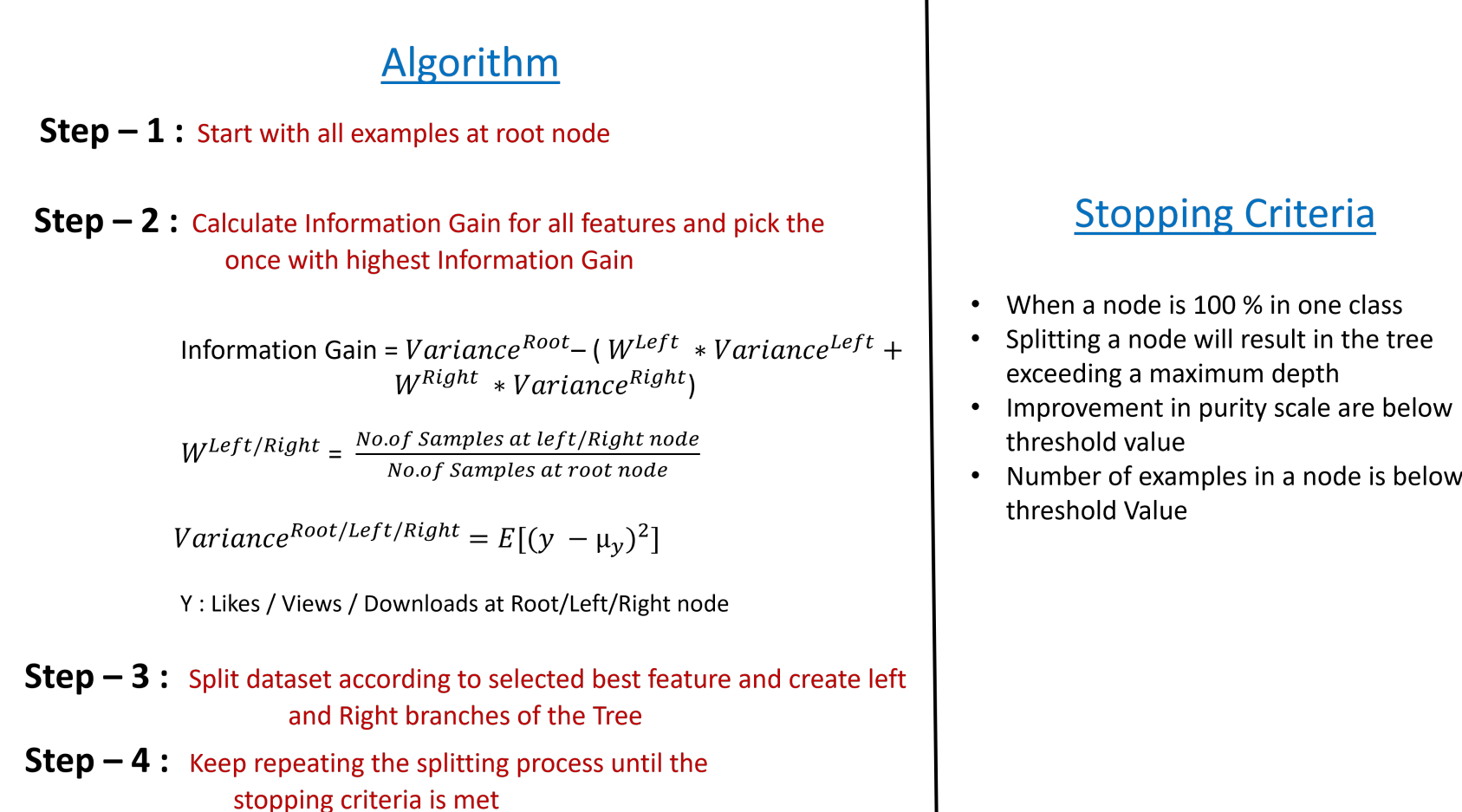
In present world, almost every person irrespective of their age have been engaging themselves in social media. So if we want to reach to a large number of population, social media must be one of the easiest way. By analyzing social media data using machine learning techniques, marketers gain deeper insights into consumer behavior, preferences, and sentiments. Machine learning models can predict individual user preferences based on their social media interactions. So we have proposed some machine learning based mathematical models. Those models will predict the consumer behaviour with different accuracy based on the machine learning approaches that we are using.

Methodology

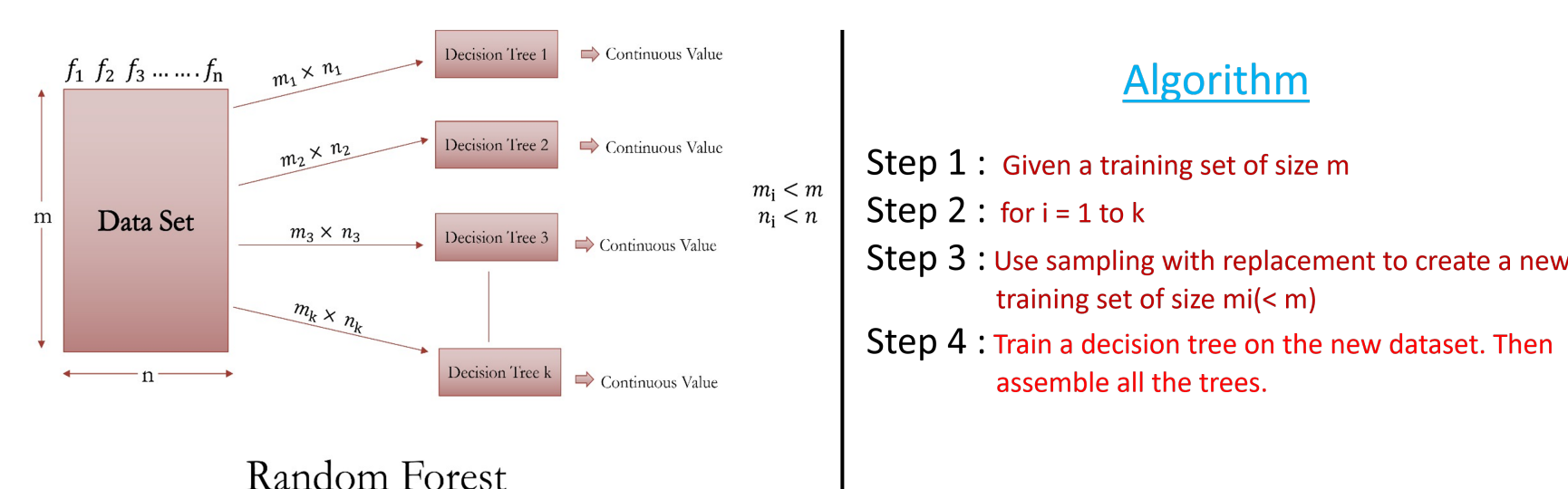
- Data Preprocessing:** We have removed outliers, errors from the data. After cleaning to do the predictions of consumer behaviour, we have used the following machine learning methods.
- Linear and Polynomial Regression:** The formula for prediction using linear regression is $\hat{y} = w.x + b$, where \hat{y} is predicted value of consumer behaviour. To find the parameter we will minimize the cost function or mean squared error $J(b, w) = \frac{1}{2n} \sum_{i=1}^n (y_i - (b + w.x_i))^2$
- The prediction for behaviour by polynomial regression is $\hat{y}(x) = b + w_1x + w_2x^2 + \dots + w_nx^n$, and we have calculated the parameters same as linear regression.



- Decision Tree:** It does not have any parameter. It has a tree like structure, that consists of a root node, branches, internal nodes and leaf nodes. We start with the root node. Then Calculate information gain for all possible features and pick the one with highest information gain. We finally calculate the error between actual and predicted value by mean squared error.



- Random Forest:** Why? We will deal with overfitting of decision tree & combine output of multiple decision trees and get a single result. We will take a subset m_i from total feature m , & train those new $m_i (< m)$ dataset. The average of the outputs of the newly built decision trees will be the final prediction.



- XGBoost:** It is an optimised distributed gradient boosting library. In this method we calculate similarity gain & similarity score while determining root and leaf node. We define residual by difference between observed value & predicted value.

$$\text{similarity score (SS)} = \frac{(\text{sum of residuals})^2}{\text{number of residuals} + \lambda}$$
$$\text{similarity gain} = (SS_L + SS_R) - SS_{root}$$

Methodology(continue)

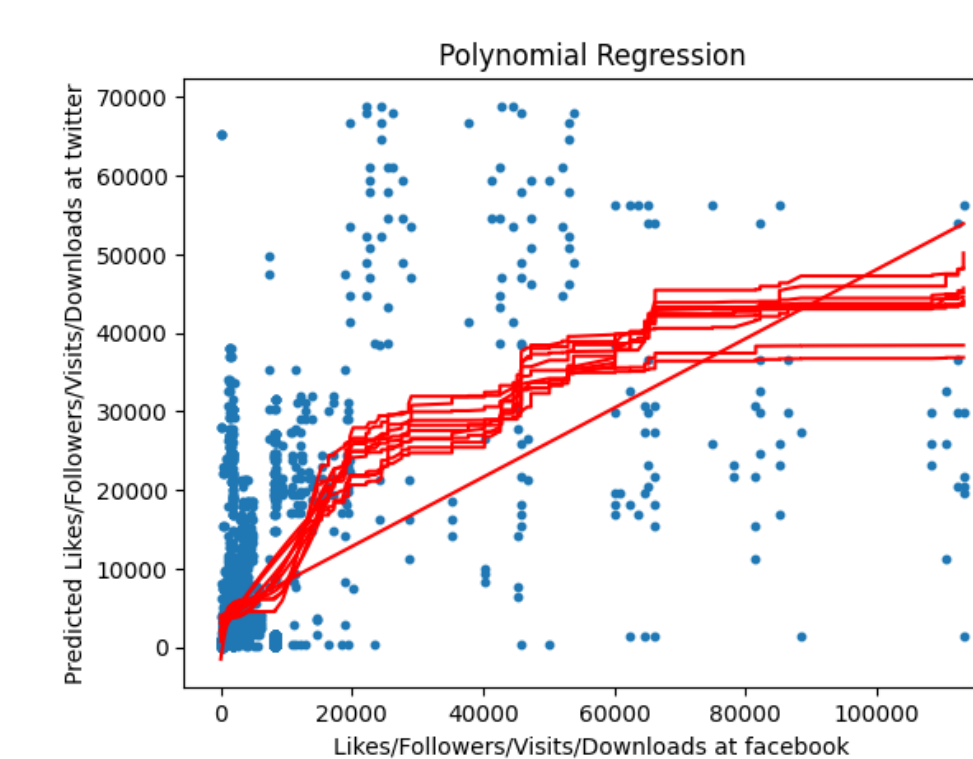
- Input:** The training set $\{(x_i, y_i)\}_i^N$, a differentiable loss function $L(y, F(x))$, number of iterations M .
- Step1:** Initialize $f_0(x) = \argmin_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$
- Step2:** For $m = 1$ to M :
 - (a) For $i = 1, 2, \dots, N$ compute $r_{im} = -\left[\frac{\Delta L(y_i, f(x))}{\Delta f(x_i)}\right]_{f=f_{m-1}}$
 - (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \dots, J_m$
 - (c) For $j = 1, 2, \dots, J_m$ compute $\gamma_{jm} = \argmin_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$

- We use one hot encoding of the dataset features for last three algorithm.

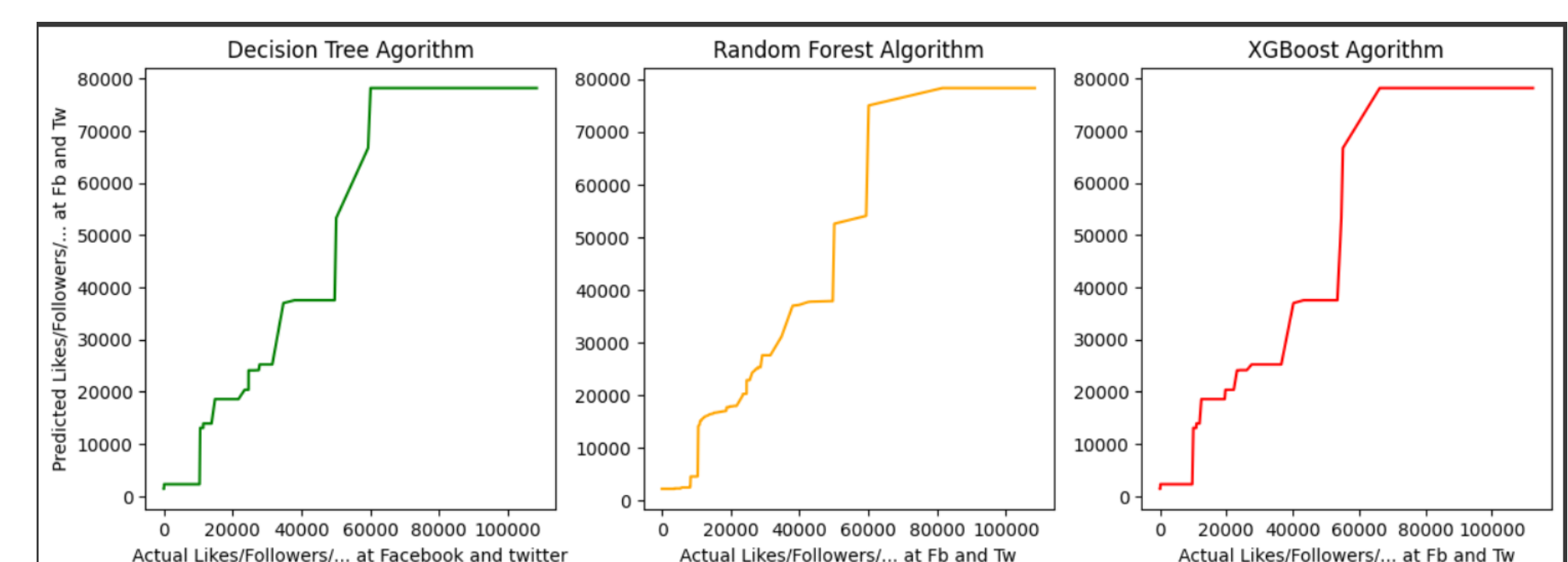
Results & Discussion

After applying above algorithms in Python we are getting the following results, where in all the graphs X axis represents the actual activity of customer in some platform, whereas Y axis represents the predictions for the same. For codes, click on https://github.com/pritamraymng/Project_2nd_sem

- Linear & polynomial Regression:**



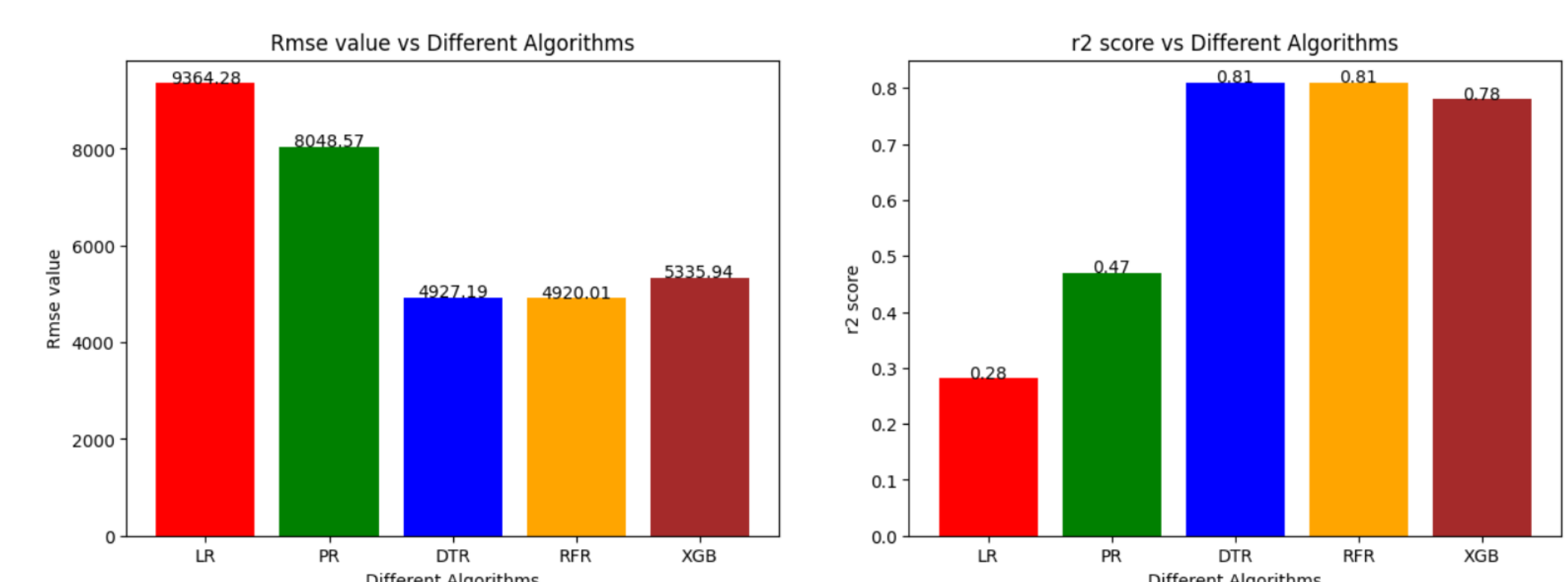
- We are getting the accuracy in consumer behaviour prediction for linear as 28%, for ten degree polynomial accuracy is 47%, & due to overfitting of curve in eleven degree polynomial accuracy is 44%.
- Decision Tree, Random Forest & XGBoost:**



- We are getting the accuracy of prediction in decision tree as 80.89%, in random forest as 81%, and in XGBoost as 78%.

Conclusion

- We can see that random forest algorithm gives the best accuracy in predicting consumer activity as 81%. So, we can select random forest algorithm as best algorithm. Here are some bar graphs.
- The first graph gives us the comparison between root means square & different algorithm.



- The second algorithm gives us the comparison between R2 score & different algorithm.

References

- Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics, Kiran Chaudhary, Mansaf Alam, Mabrook S. Al-Rakhani and Abdu Gumaei, Springer Open, 2021.
- A machine learning-based approach to enhancing social media marketing, Senthil Arasu, B. Jonath Backia Seelan, N. Thamaraiselvan, Department of Management Studies, National Institute of Technology, Tiruchirappalli, India.
- XGBoost: A Scalable Tree Boosting System, Tianqi Chen, University of Washington, tqchen@cs.washington.edu